Q&A session for Midterm

Zhaohu(Jonathan) Fan

06/15/2021

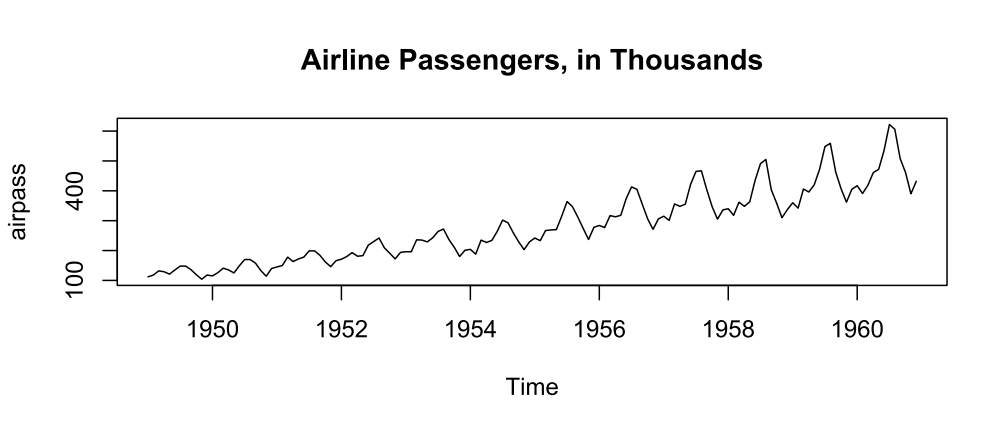
- Review of topics covered:
  - Chapter 1
    - Visualizing Time Series Data in R
  - Chapter 2
  - Chapter 3 (Part I)
- Exam

Chapter 1

Time series data

- Time series data form an ordered sequence of numbers, corresponding to an object like quantities, prices,counts, observed at or over a particular point **in time**.
- The "Airline Passengers" data set is an example of **time series data**.

```
plot(airpass)
title(main="Airline Passengers, in Thousands")
```



Time series patterns

- Describing a time series: trend, seasonality, cycles, changing variance, unusual features.
  - **Trend**: pattern exists when there is a long-term increase or decrease in the data.
  - **Seasonal** : pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week).
  - **Cyclic** : pattern exists when data exhibit rises and falls that are *not of fixed period* (duration usually of at least 2 years).

Seasonal or cyclic?

Differences between seasonal and cyclic patterns:

- seasonal pattern constant length; cyclic pattern variable length
- average length of cycle longer than length of seasonal pattern
- magnitude of cycle more variable than magnitude of seasonal pattern

The timing of peaks and troughs is predictable with seasonal data, but unpredictable in the long term with cyclic data.

Visualizing time series data in R

- Visualization is good practice to be able to understand the properties of the data.
    - Most time series coming from official data sources provide recordings at a regularly spaced set of such time points, such as every day, week, month, quarter, or year; this interval is called the **frequency** of the time series.
- A time series is stored in a `ts` object in R:
    - `ts` objects and `ts` function

For observations that are more frequent than once per year, add a `frequency` argument.

E.g., monthly data stored as a numerical vector `z`:

```
y ← ts(z, frequency=12, start=c(2003, 1))
```

Chapter 2

Correlation

- Correlation: measure of linear relationships (**a measure of the direction and strength of the relationship between two variables**)
- We use Pearson's r as a measure of the linear relationship between two quantitative variables. In a sample, we use the symbol r . In a population, we use the Greek letter ("rho"). Pearson's r can easily be computed using R.
- The correlation coefficient, rho, measures linear relationships: Ranges over [-1, +1]
    - A value of +1 indicates a perfect positive (upward sloping) linear relationship between the two variables.
    - A value of -1 indicates a perfect negative (downward sloping) linear relationship between the two variables.
    - A value of zero indicates no linear relationship between the two variables.

Interpret correlation coefficient

- Correlation coefficient is comprised between -1 and 1:
  - -0.86 indicates a strong negative correlation : this means that every time x increases, y decreases.
  - 0 means that there is no association between the two variables (x and y).
  - 0.87 indicates a strong positive correlation: this means that y increases with x .

Note: The closer r is to 0 the weaker the relationship and the closer to +1 or -1 the stronger the relationship (e.g., r=-0.98 is a stronger relationship than r=+0.78); the sign of the correlation provides direction only.

Interpret correlation coefficient

- To determine whether the correlation between variables is significant, compare the p-value to your significance level. Usually, a significance level (denoted as α or alpha) of 0.05 works well.

- An α of 0.05 indicates that the risk of concluding that a correlation exists—when, actually, no correlation exists—is 5%. The p-value tells you whether the correlation coefficient is significantly different from 0. (**A coefficient of 0 indicates that there is no linear relationship.**)

    - **P-value ≤ α**: The correlation is statistically significant. If the p-value is less than or equal to the significance level, then you can conclude that the correlation is different from 0.
    - **P-value > α**: The correlation is not statistically significant. If the p-value is greater than the significance level, then you cannot conclude that the correlation is different from 0.

Chapter 3 (Part I)

Residual diagnostics

- Residuals in forecasting: difference between observed value and its fitted value: $e_t = y_t - \hat{y}_{t|t-1}$.
- Assumptions:
  - (1) $\{e_t\}$ are **uncorrelated**. If they aren't, then information left in residuals that should be used in computing forecasts.
  - (2) $\{e_t\}$ have **mean zero**. If they don't, then forecasts are biased.
- Useful properties (for prediction intervals):
  - (1) $\{e_t\}$ have **constant variance**.
  - (2) $\{e_t\}$ are **normally distributed**.

Autocorrelation (ACF)

Covariance and **correlation**: measure extent of **linear relationship** between two variables (Y and X).

Autocovariance and **autocorrelation**: measure linear relationship between **lagged values** of a time series y.

We measure the relationship between:

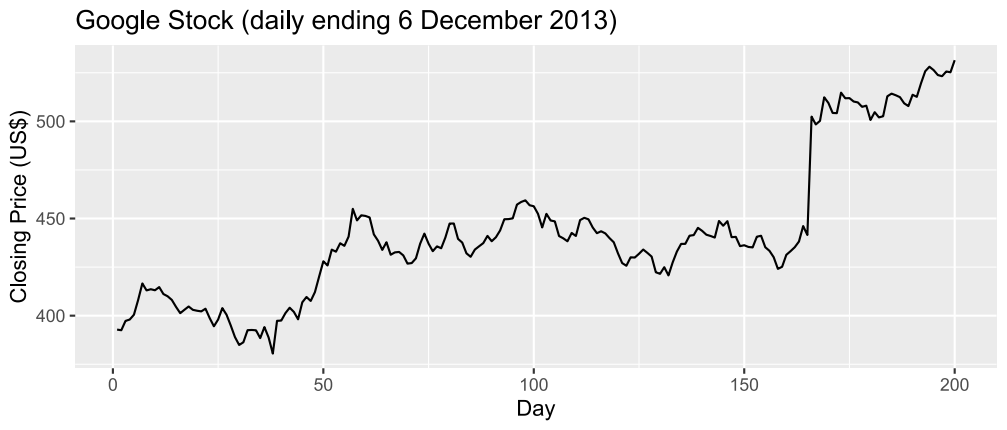
- y_t and y_{t-1}
- y_t and y_{t-2}
- y_t and y_{t-3}
- etc.

ACF of residuals

- We assume that the residuals are white noise (**uncorrelated**, mean zero, constant variance). If they aren't, then there is information left in the residuals that should be used in computing forecasts.
- So a standard residual diagnostic is to check the ACF of the residuals of a forecasting model.
- We expect these to look like white noise.

Data Example: Google stock price

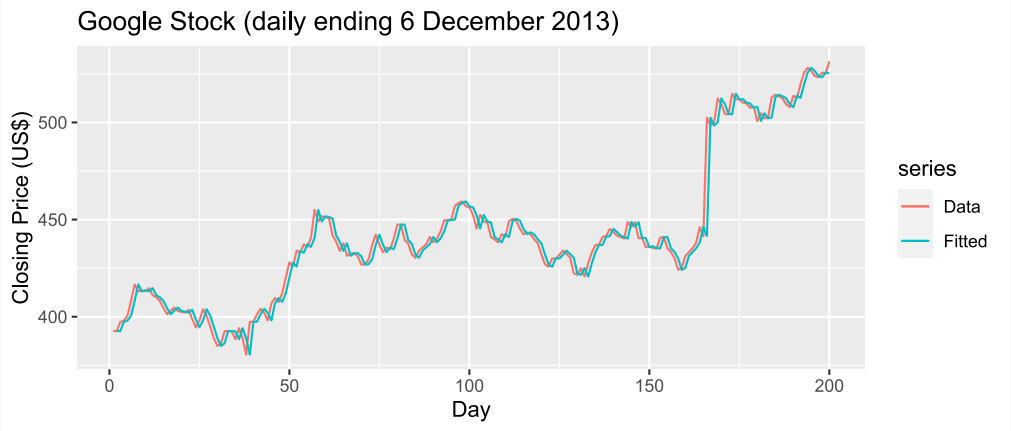
```
autoplot(goog200) +
  xlab("Day") + ylab("Closing Price (US$)") +
  ggtitle("Google Stock (daily ending 6 December 2013)")
```



Data Example (cont'd)

- Naïve model

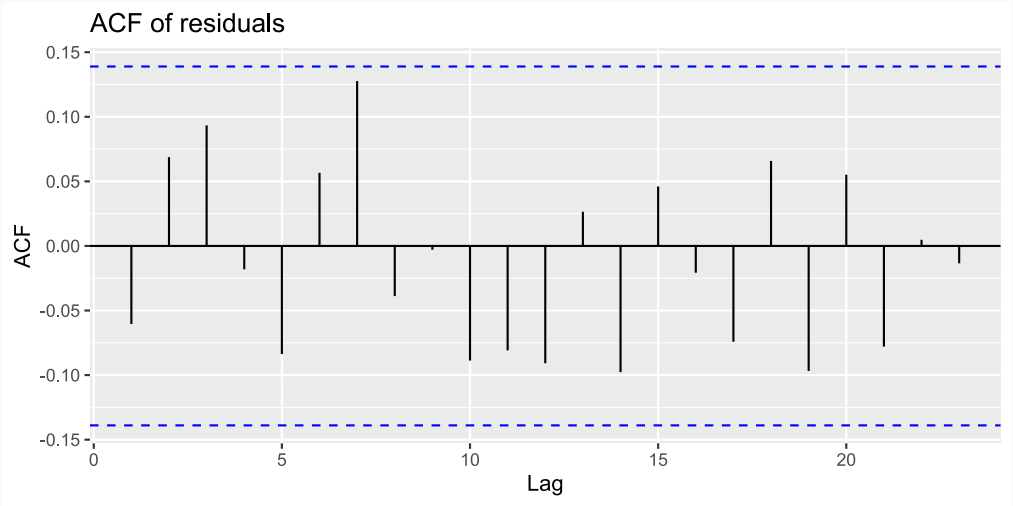
```
fits ← fitted(naive(goog200))
autoplot(goog200, series="Data") +
  autolayer(fits, series="Fitted") +
  xlab("Day") + ylab("Closing Price (US$)") +
  ggtitle("Google Stock (daily ending 6 December 2013)")
```



Data Example (cont'd)

- check if the residuals are white noise (**uncorrelated?**).
- We expect each autocorrelation to be close to zero.

```
res ← residuals(naive(goog200))
ggAcf(res) + ggtitle("ACF of residuals")
```

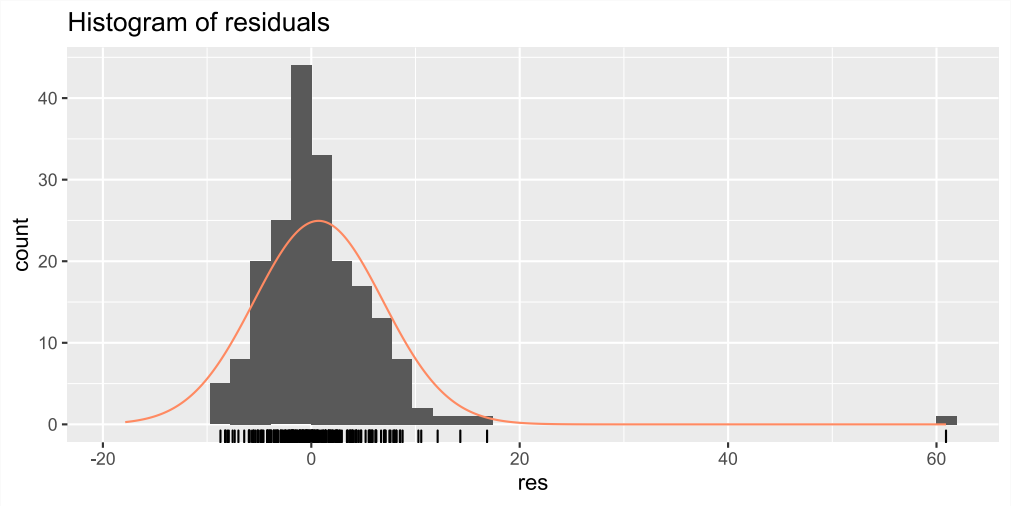


- There is no significant correlation in the residuals series.The vertical lines (lags) don't pass through horizontal blue lines. This means the autocorrelations are not (statistically) significantly different from zero.

Data Example (cont'd)

- check if the residuals are white noise (**normally distributed+ mean zero?**).

```
gghistogram(res, add.normal=TRUE) +
  ggtitle("Histogram of residuals")
```

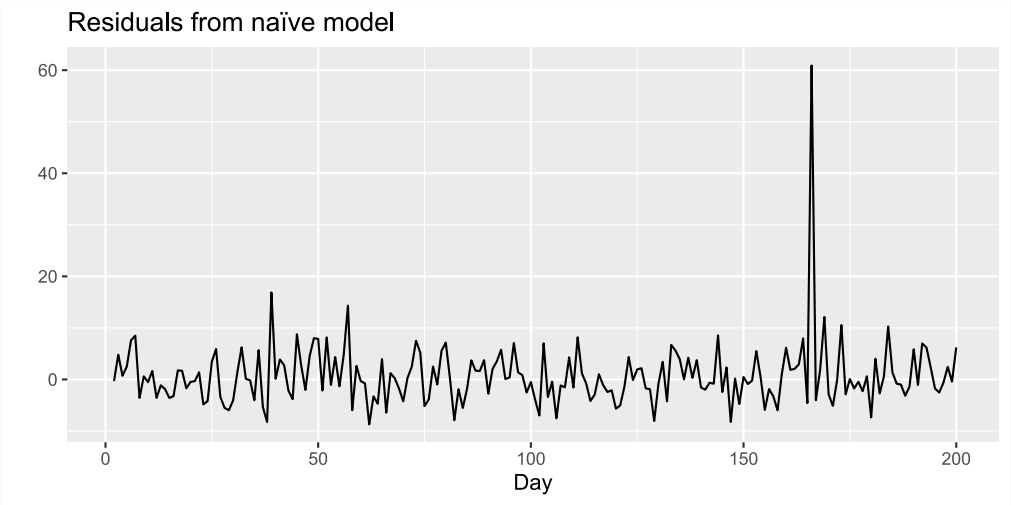


- The mean of the residuals is close to zero.
- The histogram suggests that the residuals may not be normal. The right tail seems a little too long, even when we ignore the outlier.

Data Example (cont'd)

- check if the residuals are white noise (**constant variance?**).

```
autoplot(res) + xlab("Day") + ylab("") +
  ggtitle("Residuals from naïve model")
```



- The time plot of the residuals shows that the variation of the residuals stays much the same across the historical data, apart from the one outlier, and therefore the residual variance can be treated as constant.

Data Example (cont'd): Conclusion

- Forecasts from this method (Naïve model) will probably be quite good,
- but prediction intervals that are computed assuming a normal distribution may be inaccurate.

Ljung-Box test

- In general, the Ljung-Box test (a test of autocorrelation of the residuals) is defined as:
  - H_0 (null hypothesis): the autocorrelations come from a white noise series.
  - H_a (alternative hypothesis): the autocorrelations do not come from a white noise series.

- Ljung-Box test is based on

  - $Q^* = T(T+2)\sum_{k=1}^{h}(T-k)^{-1}r_k^2$ where h is max lag being considered and T is number of observations.
  - large values of Q^* suggest that the autocorrelations do not come from a white noise series.

- By convention: $h = 10$ for non-seasonal data, $h = 2m$ for seasonal data, where m is the period of seasonality.

- Better performance, especially in small samples.

Ljung-Box test (cont'd)

- If data are white noise, Q^* has χ^2 distribution with $(h-K)$ degrees of freedom where $K =$ no. parameters in model.
- For the Google stock price example, the naïve model has no parameters, so $K = 0$

```
# lag=h and fitdf=K
# Total lags used: 10
Box.test(res, lag=10, fitdf=0, type="Lj")
```

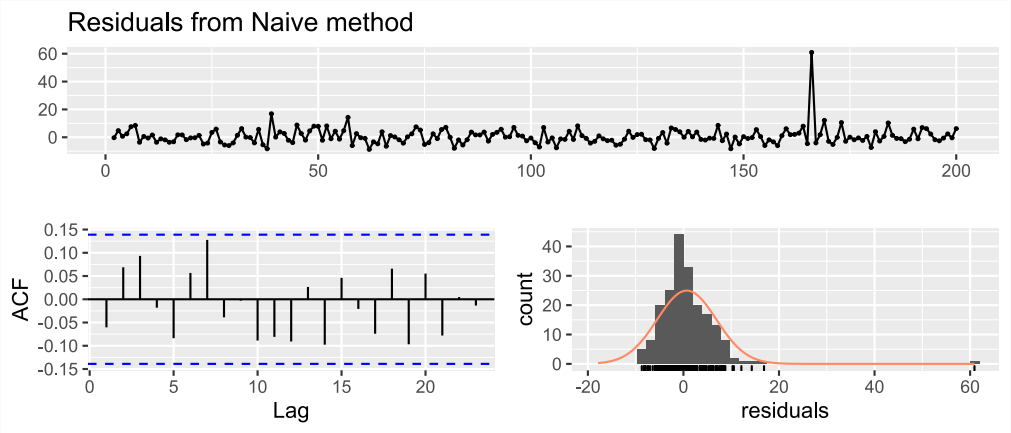
```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 11.031, df = 10, p-value = 0.3551
```

- **The results are not statistically significant (i.e., the p-value>0.05). We fail to reject the null hypothesis H_0. Thus, we can conclude that the residuals are not distinguishable from a white noise series**.

Ljung-Box test (cont'd)

- Test if the residuals are white noise.
- `checkresiduals` function produces 1) a time plot of residuals, 2) a ACF of the residuals, 3) a histogram of the residuals (with an overlaid normal distribution for comparison), and 4) a Ljung-Box test.

```
checkresiduals(naive(goog200))
```

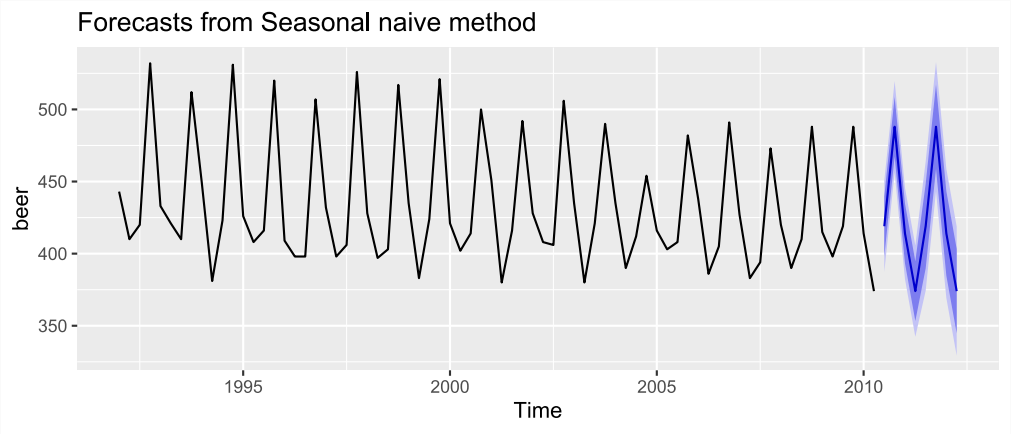



```
##
##  Ljung-Box test
##
## data:  Residuals from Naive method
## Q* = 11.031, df = 10, p-value = 0.3551
```

Your turn

Compute seasonal naïve forecasts for quarterly Australian beer production from 1992.

```
beer ← window(ausbeer, start=1992)
fc ← snaive(beer)
autoplot(fc)
```



Test if the residuals are white noise.

```
checkresiduals(fc)
```

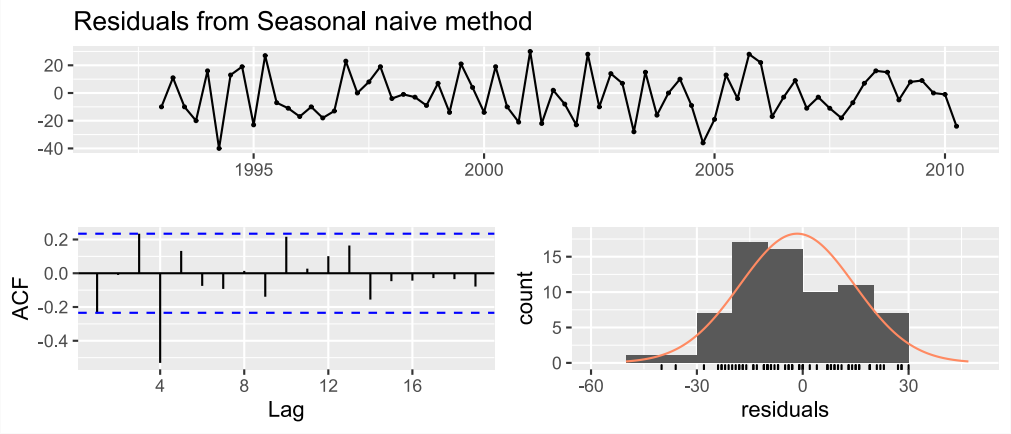
What do you conclude?

Sample Solutions

Test if the residuals are white noise.

- **The results are statistically significant (i.e., the p -value< 0.05). We reject the null hypothesis H_0. Thus, we can conclude that the residuals are distinguishable from a white noise series**.

```
checkresiduals(fc)
```




```
##
##     Ljung-Box test
##
## data:  Residuals from Seasonal naive method
## Q* = 32.269, df = 8, p-value = 8.336e-05
```